RCT-G2

POSTAL GEOCODING SYSTEMS AND SMALL AREAS

Administrative Data Applications

Louise Radley,
Nelson Kopustas,
Administrative Data Development Staff,
Statistics Canada,

June, 1981

TABLE OF CONTENTS

		Page
I.	INTRODUCTION	1
II.	ANALYSIS AND CODING OF THE MAILING ADDRESS	4
	- THE MAILING ADDRESS AND PLACE OF RESIDENCE	7
	- CODING THE MAILING ADDRESS	8
	- OBJECTIVES	9
III.	APPLICATIONS: CENSUS DIVISIONS	10
	A. Preliminary Matching Procedure and Results	12
	B. Final Matching Procedure and Results	16
IV.	APPLICATIONS: CENSUS TRACTS	21
	A. Calgary	22
	B. Ottawa-Hull	30
	C. Summary of Census Tract Pilot Studies	33
V.	CONCLUDING REMARKS	34
	A. Summary of Findings	34
	B. Some Proposed Solutions	36
	C. Concluding Overview	37
BIBLIOGRAPHY		40
APPENDICES		41

LIST OF TABLES

		Page
TABLE 1:	Three Digit Locality Code Match.....	15
TABLE 2:	Summary Table 1975-1978 for 5-digit Locality Code.....	19
TABLE 3:	Results of Census Tract Coding for Calgary.....	25
TABLE 4:	Examination of Records with Blank Census Tracts for Calgary When the Postal Code is Valid.....	27
TABLE 5:	Summary of Results from Two Pilot Projects.....	31

LIST OF FIGURES

		Page
FIGURE 1:	Rural Post Office Service Areas in Eastern Ontario.....	6
FIGURE 2:	Procedure for Consistency Check of Census Division Coding.....	14
FIGURE 3:	Creation of Census Tract Coded File for Calgary for 1976.....	23

I. INTRODUCTION

Social programs have grown dramatically in the last three to four decades. As these programs have evolved, the demand for more frequent small area social data has also grown. Both producers and users of social data have endeavoured to close the gap between what is available at the small area level, and what is needed at the small area level.

Statistics Canada, in considerable measure, has produced social data through social or household surveys and the census of population program. But neither of these has satisfied the unmet needs. The survey data cannot be dissaggregated to sufficiently small areas for policy, evaluative or other purposes. The census, while meeting the small area criterion, is neither frequent nor timely enough.

Increasingly, administrative records were proposed by users as a means of meeting some of the small area data needs. Administrative records, or so it was argued, could be disaggregated to small areas and could be available annually. In response to these and other forces, in 1979 Statistics Canada established the Administrative Data Development Staff (ADDS) to assess the social data potential of administrative records.

Given this broad mandate, the first task was the identification of those administrative records that could provide social data for all provinces and territories. Four sets of records met this first test: the personal income tax (T1) records, the family allowance records, the Old Age

Security records, and the unemployment insurance beneficiary records. Of these four sets of records, the one containing the largest array of potential social data was the T1 record set. Thus, during the first two years of the development work of ADDS' almost complete attention was allocated to the T1 file.

Two additional questions were posed: First, what kind of data can be potentially produced from the T1 file? and secondly, how is a small area to be defined? In answer to the first question, it was concluded that the T1 file could enable the production of data on migration, income and the labour force with cross-tabulations by age, sex and marital status.

With regard to the second question, it was apparent that two approaches could be identified. First, small areas could be defined in ways consistent with the standard geographical coding system of Statistics Canada. Or, alternatively, small areas could be defined in an open-ended way to enable the assignment of individual records to an almost infinite array of user-specified areas.

In spite of the apparent dilemma, it was deemed essential to assess the basic quality, comparability and consistency of administrative tax data to similar data derived from the Census of Population. Clearly, the data derived from administrative records would be of little value if at least low grade comparability to census data could not be obtained for small areas. It was concluded that if the geographical information on

administrative records could not be reliably aggregated into small areas, then there would be no future for the production of any kind of administrative social data. Therefore, from the very beginning, a substantial investment in geographical delineation became the cornerstone for the subsequent development of social data from administrative records.

The remainder of this paper has been designed to review the mailing address of the T1 file; the conversion of this mailing address into counties or census divisions; and a discussion of two pilot studies for the conversion of mailing address information into census tracts.

II. ANALYSIS AND CODING OF THE MAILING ADDRESS

All of the geographical information of the T1 file is contained in or derived from the mailing address. And in general, the mailing address can be subdivided into four classess of mail delivery:

a. door-to-door mail delivery,

b. suburban service,

c. rural route, and

d. post office box or general delivery.

With door-to-door delivery, the mailing address identifies a geographical location rather precisely in urban areas. When the postal code is also present, an efficient six character alpha-numeric code is also present and identifies a block face.

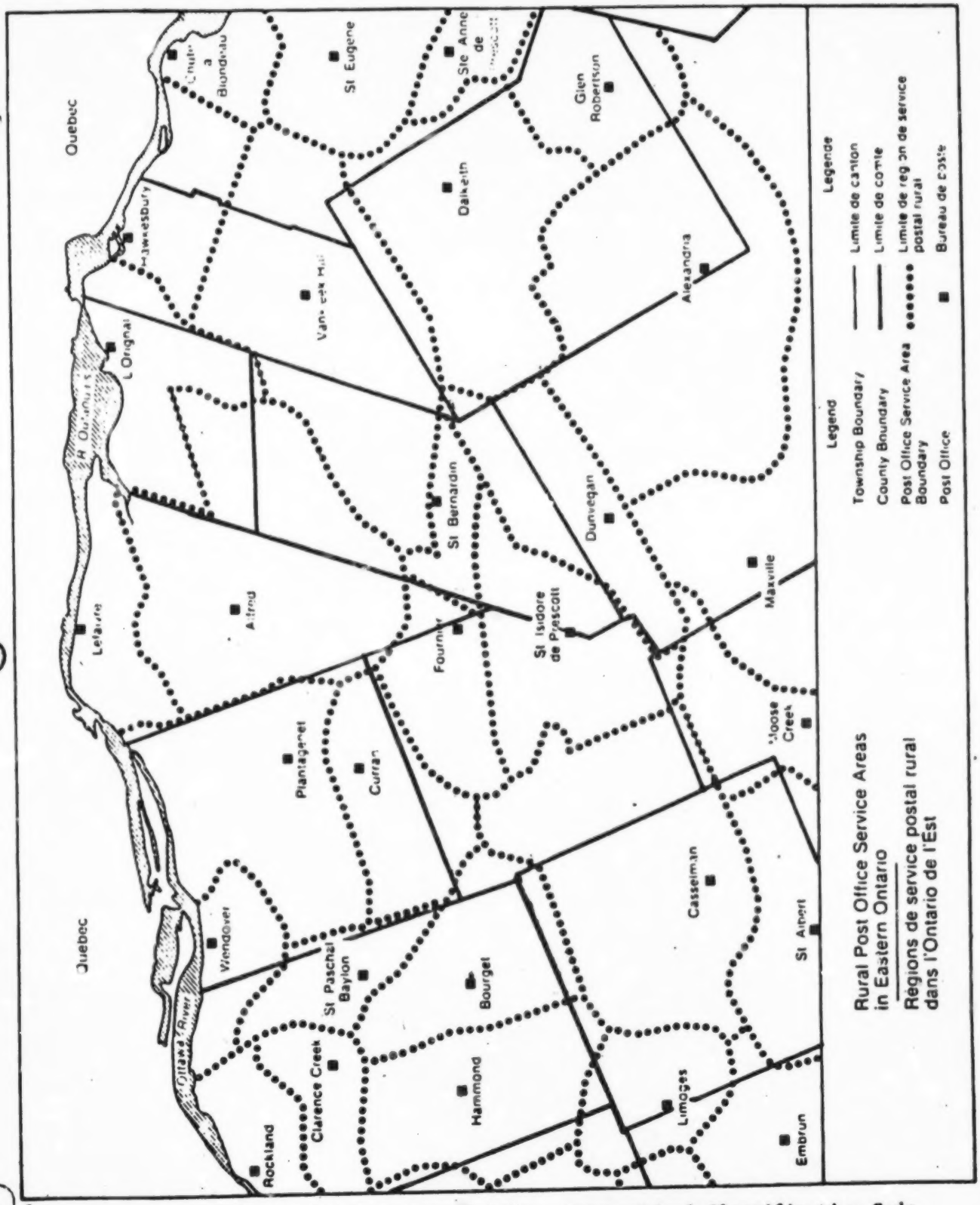
With suburban service, typical of new suburban developments, the post office delivers mail to designated places where residents can pick up their mail from individually padlocked boxes. In general, the geographical information derived from the mailing addresses of this mail distribution system is imprecise and does not enable the allocation of unique geographical codes to individual records.

With rural delivery, the mailing address specifies a route number from a specific post office location. Although the residence does not coincide with the post office location, it is impossible to assign the

address information to any other geographical location. Therefore, all geographical assignments for rural routes are the same for all rural route mailing addresses with the same place name. For example FIGURE 1 includes the locations of post offices for an area within and bordering on one census division. It can be noted in this figure that some of the postal service areas are wholly within the census division while others have only parts of their service areas within the census division. In the boundary cases when the post office is located within the census division, all taxfilers from that post office location are included in the census division, even those who reside outside of the census division. Similarly, other boundary post offices are located outside of the census division but have service areas within the census division. Clearly, there is considerable fuzziness around the boundaries of rural census divisions. Nevertheless, the evidence seems to indicate that the misallocations approximately cancel each other out.

With post office box number and general delivery, the customer picks up his/her mail at the post office. In both cases, the location of the post office can be uniquely identified, but the geographical location of the customer cannot.

FIGURE 1



Source: Standards Division, Postal Code - Geographical Classification Code Conversion File User's Guide, March 1979.

THE MAILING ADDRESS AND PLACE OF RESIDENCE

As indicated in the preceding section, the mailing address can differ from the place of residence. Insofar as data derived from censuses and surveys are based on residential locations, administrative tax data can diverge from this concept when the mailing address does not identify the same geographical location as the residential address.

For example, if the mailing address is a street address, it is reasonable to assume that the mailing address and residential addresses are identical most of the time. If, however, the address refers to an office or a store, a discrepancy can be identified. In other cases, the mailing address may be that of a third party such as a taxfiler's accountant or lawyer. And for the other three cases of mail delivery, it is not possible to identify very precise geographical locations.

Given the array of potential geographical predicaments that can be easily identified, the reason for investigating the conversion of mailing address information into standard geographical systems should be apparent. It would seem foolhardy to begin tabulating data without conducting a careful examination of the geographical implications of using the mailing address in the production of statistical information for small areas.

CODING THE MAILING ADDRESS

The mailing address is lengthy and unstructured. It is not suited to the computer processing of large files such as the T1 with 12-14 million records. Reliance, therefore, has been placed on two summary codes, the postal code and the locality code.

The postal code is a six character code that groups mailing addresses. For example, addresses are grouped by a range of street addresses in urban areas with door-to-door mail delivery; a range of post office boxes; or a group of rural routes with the same post office. The postal code, from this perspective, contains almost as much information in six characters as the mailing address does in sixty characters.

The locality code is a five-digit code that was created and maintained by Revenue Canada-Taxation. The first three digits were designated to represent a census division or the name of a metropolitan area. The last two digits summarize the name of a municipality. For example, the locality code 21711 identifies census division 217 (Russell County) and municipality 11 (Cumberland Township). For the ease of data processing, therefore, the postal code and locality code have been used extensively in lieu of the mailing address for assigning standard geographical codes to individual tax records.

OBJECTIVES

With the overview provided in this section, it is now possible to summarize the objectives identified for this paper:

a. To describe the methodology for converting both postal codes and locality codes to SGC codes.

b. To assess the consistency of the two methods for assigning SGC codes to the same individual records.

c. To test the postal code to census tract conversion process for two metropolitan areas.

d. To provide a general assessment of the extent to which mailing addresses approximate place of residence.

III. APPLICATIONS: CENSUS DIVISIONS

The first two objectives noted above will be addressed explicitly in this section of the paper -- converting both postal codes and locality codes to SGC codes, and then assessing the consistency of these two methods of assignment.

To convert postal codes and locality codes to SGC codes, three inputs were required:

(1) the 1976 T1 file containing records with both the locality code and postal code present,

(2) a table relating postal codes to SGC census division codes (the Standards Division conversion file[1]), and

(3) a table relating locality codes to SGC census division codes (created by ADDS).

(1) The T1 file contains one record for each person who filed a tax return (i.e., taxfiler). In 1976, there were 12,339,280 records on the file and of these 7.4%(911,056) did not contain postal codes. For the remaining 92.6%(11,428,224), both the locality code and postal code were present. From the records containing both codes, a sample of approximately one million records was randomly selected for processing.

(2) The conversion file created by Standards Division contains 593,000 records, one for every postal code in Canada. Each postal code

[1] Statistics Canada. Standards Division, Postal Code -- Geographical Classification Code Conversion File User's Guide, Ottawa, March 1979.

on this file has been assigned the appropriate code for a number of geographic coding (i.e., SGC code, Census Enumeration Area code, Census Tract code, and Census Agglomeration code).

(3) Three digit locality codes for the 1976 tax year were assigned 1976 SGC census division codes on the basis of census division names and city names. An urban area was assumed to be wholly contained within the census division it was listed under (e.g., St. John's, Newfoundland, was assumed to be totally within Census Division 1). Since some 3-digit locality codes are combinations of SGC defined census divisions, these were excluded from the processing. Examples of these combined census division codes are listed in APPENDIX 1. Therefore, the resulting locality code to census division table had a one-to-one correspondence between census division and 3-digit locality codes. This table included approximately 90% of the census divisions of Canada. (By accepting those locality codes that are combinations of SGC defined census divisions, one could effectively include all census divisions. However, for the purpose of determining consistency, these were excluded.)

A. Preliminary Matching Procedure and Results

The general procedure is shown in FIGURE 2. The 1976 T1 file was first matched to the postal code - census division code table and a census division code (CDp) was assigned to each record on the basis of its postal code. Records containing a postal code not found in the table were dropped and a frequency by postal code was printed for examination (i.e., Report 1 of FIGURE 2).

Those records with a census division code assigned from the postal code (CDp) were then matched to the locality code - census division code table and a second census division code was assigned on the basis of locality code (CD_L). Records containing a locality code not found in the table were dropped and a frequency by locality code was printed (i.e., Report 2).

The records with two census division codes (i.e., CDp and CD_L) were then compared. A file of records with $CDp=CD_L$ was created and the frequency of these records by locality code was printed (i.e., Report 3). A second file of records with $CDp \neq CD_L$ was created and a frequency by postal code within the locality code for each (CDp, CD_L) combination was printed (i.e., Report 4).

The first results obtained from the match on the 3-digit locality code are shown in TABLE 1. An examination of Report 1 turned up postal codes for rural routes no longer in existence and

these were added to the postal code table for future use. The locality code table was also updated for future use as a result of locating invalid locality codes from frequencies printed in Report 2. An examination of Report 4 revealed RC/T miscoding; that is, many cases appeared where place names were assigned incorrect locality codes. As mentioned earlier, there are 3-digit locality codes that are combinations of SGC defined census divisions (see APPENDIX 1) and these accounted for 11.32% of the total records. (Examples of tabulations obtained from Reports 1, 2, 3 and 4 are found in APPENDIX 2.)

FIGURE 2

Procedure for Consistency Check of Census Division Coding

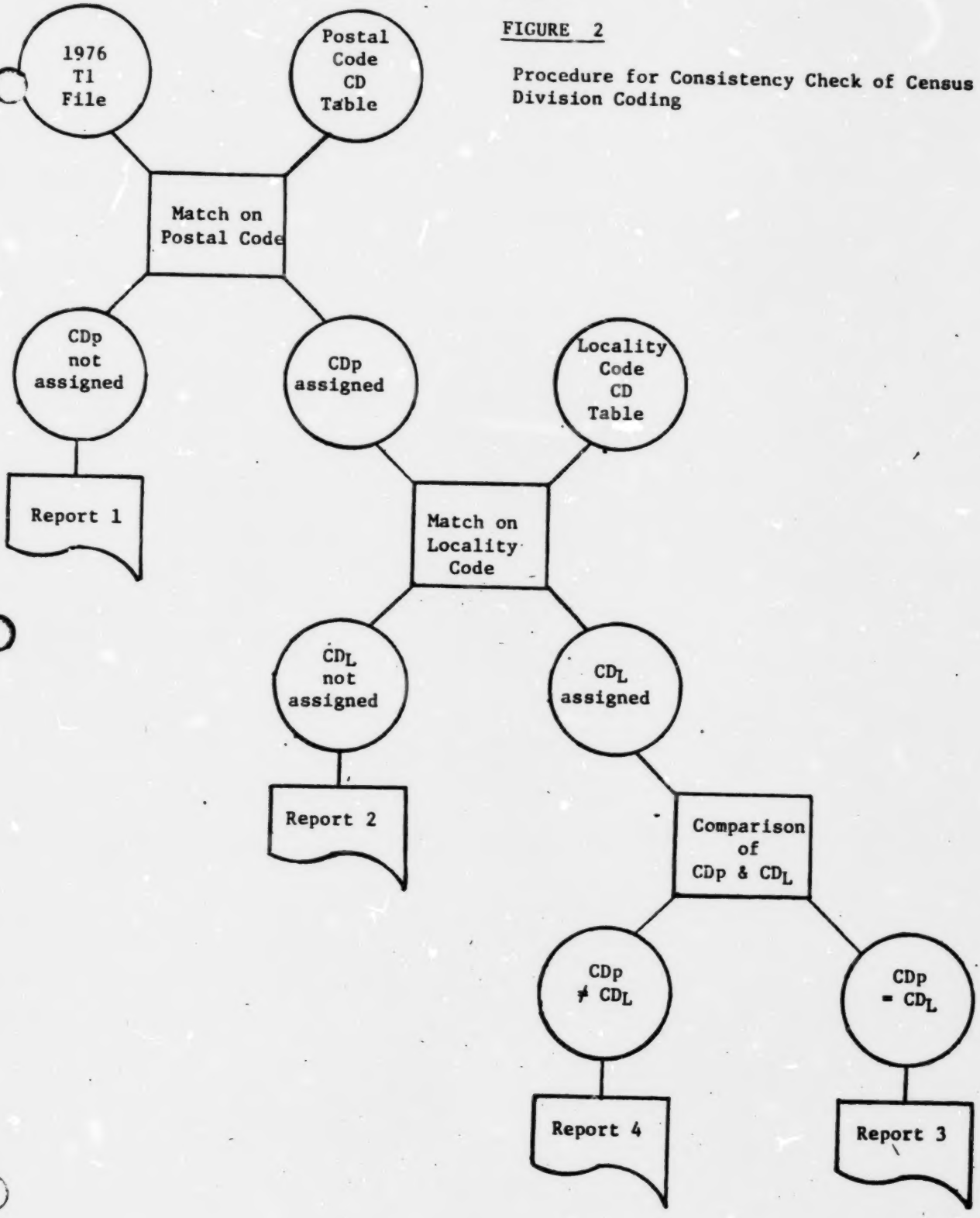


TABLE 1

Three Digit Locality Code Match

	No.	%
T1 file records	988,518	100.00
Locality code in combined census divisions	111,890	11.32
CDp not assigned (Report 1)	7,868	0.80
CD_L not assigned (Report 2)	101	0.00
CDp not equal to CD_L (Report 4)	24,069	2.43
CDp equal to CD_L (Report 3)	844,590	85.44

With an inconsistency rate of approximately 14 percent, it was apparent that a problem existed in the conversion process. Subsequent investigation indicated that the divergencies could be attributed to the 3-digit locality code. This led, therefore, to the construction of a five-digit locality code - census division code table.

Using the information in Reports 1, 2 and 4, the postal code - census division code table and the locality code - census division code table, both conversion tables were modified and the process in FIGURE 2 was repeated. The objective was to minimize the number of entries in these three reports.

B. Final Matching Procedure and Results

Using the RC-T locality code manual[2], each place name was assigned a 1976 SGC code using the SGC manual[3]. When a place name could not be found in the SGC manual, gazeteers and maps were used to assign a census division code where possible. Then a 5-digit locality code - census division code conversion table was constructed for the machine conversion process. And since the existence of combined census divisions disappears when 5-digit locality codes are used, all census divisions were included in the final matching procedure.

---

[2] "Post Office Communities and Place Names with their Respective District Office Codes and Locality Codes for All Canada by Province in Numerical Sequence of Locality Codes", Ottawa, August 1975.

[3] Statistics Canada. Standards Division. Standard Geographical Classification Manual, June 1976, Volume III - Alphabetic Index by Province. Catalogue No.12-556, Ottawa.

The objective of this final matching procedure was to construct an improved locality code - census division conversion table using 5-digit codes. The 5-digit code is better than the 3-digit code for two reasons.

First, the locality code structure may not have been updated to reflect recent boundary changes. For example, Cumberland Township in Ontario is coded to 21711 in Russell County (3503), but should now be coded to Ottawa-Carleton Regional Municipality (3506).

Second, a taxfiler may be allocated to the wrong locality code because the place name is duplicated in the province and the incorrect one is chosen for the locality code assignment. A good example of this shows in the case of "Anjou". There is an Anjou in Berthier County (2449) and also an Anjou in Montréal (2465). All tax records with place name "Anjou" were assigned locality code "06914" which is in Berthier County. An examination of postal codes for these records revealed that approximately 90 percent should have been allocated to Montréal and 10 percent to Berthier. If the locality code "06914" is coded to census division 2449 (Berthier) 90 percent of the cases will be miscoded; if the locality code is coded to census division 2465(Montréal) 10 percent of the cases will be miscoded. The latter was chosen to minimize the error.

The process in FIGURE 2 was repeated using the 5-digit locality code - SGC census division code table with old rural postal codes incorporated in the postal code - SGC census division code table. The results for 1976 and three other years are included as TABLE 2.

For 1976, of the 0.58%(5,683) records for which an SGC census division code was not assigned from the postal code, 78.4%(4,453) were single occurences that could indicate postal code transcription errors resulting in invalid postal codes. The 1.04%(10,228) records assigned different census division codes indicate the measure of the inconsistency of the locality codes and postal codes.

TABLE 2

Summary Table 1975-1978
For 5-Digit Locality Code

	1975		1976		1977		1978	
	Number	%	Number	%	Number	%	Number	%
T1 FILE RECORDS	1,744,010	100.00	981,148	100.00	1,716,942	100.00	1,729,002	100.00
UNMATCHED POSTAL CODES	9,888	0.57	5,683	0.58	12,985	0.76	14,897	0.86
UNMATCHED LOCALITY CODES	227	0.01	103	0.01	142	0.01	442	0.03
CD MATCHED (SGC DIFF)	20,247	1.16	10,228	1.04	21,389	1.25	25,239	1.46
CD MATCHED (SGC EQUAL)	1,713,648	98.26	965,134	98.37	1,682,426	97.99	1,688,424	97.65

Note: Due to the rounding, sum of percentages may not equal 100%.

98.37 percent of the records were assigned to the same SGC census division code from the two conversion files. This indicates a high degree of consistency between the assignment reliability of each. In conclusion, therefore, when both the postal code and locality code are present and valid, the results are virtually identical.

IV. APPLICATIONS: CENSUS TRACTS

Two pilot studies were undertaken to determine the feasibility of using postal codes to allocate individual taxfiler records to census tracts. The CMAs selected were Calgary and Ottawa-Hull.

Calgary was chosen as an example of CMA's with the following characteristics:

a. rapid growth -- to assess the nature and extent of problems;

b. numbered streets -- to assess any particular requirements or difficulties in assigning census tract codes when an extensive array of streets are described by numbers rather than names; and

c. compass directions -- to assess any unique challenges arising when a compass direction is an integral part of street addresses.

Ottawa-Hull was selected for different characteristics than Calgary:

a. multiple municipalities -- to assess any special considerations arising when more than one legal municipality exists in a CMA;

b. inter-municipal streets -- to assess the census tract coding problems that arise for CMAs in which streets cross municipal boundaries and undergo name changes as well; and

c. French/English addresses -- to assess any special considerations that arise for addresses that can be formatted as either French or English.

A. Calgary

Three methodological steps were required to produce the 1976 Calgary T1 file with census tract coded individual records. The three steps were as follows:

a. the selection of all taxfiler records for the Calgary CMA from the 100 percent T1 file;

b. the creation of a postal code - census tract conversion file for Calgary; and

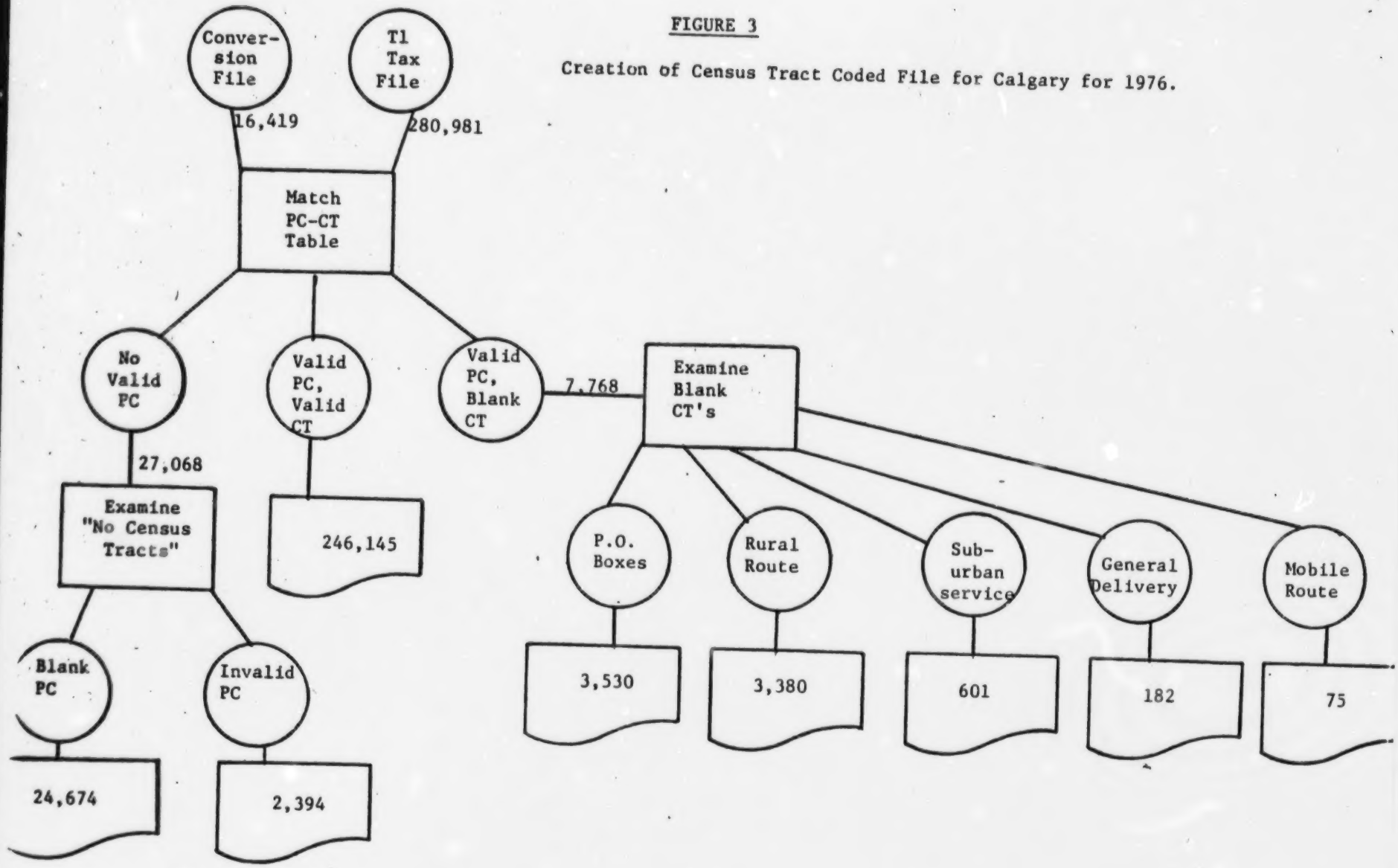
c. the matching of the above two files to assign a census tract code to each individual record in the Calgary T1 file.

These steps are illustrated in FIGURE 3.

a. In the first step, the individual records for the Calgary CMA were selected using either the locality code "42800" or the postal codes beginning with "T2", "T3" or "T1Y". This selection resulted in a file containing 280,981 individual records.

FIGURE 3

Creation of Census Tract Coded File for Calgary for 1976.



b. In the second step, the postal code - SGC conversion file of the Standards Division was used again. From this file, 16,419 postal codes for Calgary were extracted -- 90 percent or 14,729 had census tract codes. The remaining 10 percent, 1,690, were examined and two classes of cases were identified:

i. 1,444 postal codes were found that could be assigned census tract codes; and

ii. 246 postal codes for rural routes and post office boxes that could not be assigned unique census tract codes.

All but three of the 1,444 codes were manually coded to census tracts using a street map of Calgary. The remaining three street names could not be found and therefore the postal codes could not be assigned to a census tract.

c. In the third step, the 280,981 taxfilers in the Calgary T1 file were sorted by postal code and matched to the postal code - census tract conversion file. The results of this step are summarized in TABLE 3.

TABLE 3: Results of Census Tract Coding for Calgary

	Number	Percent
Valid PC, CT assigned	246,145	87.6
PC field blank	24,674	8.8
PC field invalid	2,394	0.8
Valid PC, but CT not assigned	7,768	2.8
Total	280,981	100.0

2,394 postal codes were invalid. Of these, 1,358 did not have valid Calgary postal codes. The selection of these invalid records from the 100 percent file arose through the selection on the basis of locality code "42800". Obviously, some taxfilers used a postal code for a location outside Calgary while coding their municipality name as "Calgary".

For the 27,068 addresses without a postal code, software was developed for the automatic assignment of a postal code. As a result, 16,254 additional records were postal coded and the total number of records with a valid postal code was increased to 262,399, or about 93 percent of the Calgary CMA T1 file.

The records with postal codes that could not be assigned to census tracts were examined. A summary of the results of this examination are reported in TABLE 4.

TABLE 4: Examination of Records with Blank Census Tracts for Calgary when the Postal Code Is Valid

	Number	Percent
Post Office Boxes	3,530	45.4
Rural Routes	3,380	43.5
Suburban Service	601	7.7
General Delivery	182	2.4
Mobile Route	75	1.0
Total Blank CT's	7,768	100.0

From a procedural perspective, the results were encouraging even though three problems remained:

i. some tax records did not have a postal code,

ii. the conversion file was incomplete, and

iii. not all postal codes can be assigned census tracts codes.

The first two lend themselves to resolution through the development of software or the use of clerical resources. The third appears to be intractable at this time. Although other data sources (e.g., record linkage) may prove useful in reducing its magnitude.

The census tract taxfiler counts were compared to the 1976 Census of Population data for the 84 census tracts in the Calgary CMA (see Map 1, APPENDIX 3). Two comparisons of coverage were made:

i. taxfiler counts to total 1976 population; and

ii. taxfiler counts to total 1976 population 15 years of age and older (see Table (i), APPENDIX 3).

The taxfiler universe consists of those individuals who file tax returns and have income subject to taxation (but do not necessarily pay any taxes). Since all persons in the census population do not have income subject to taxation, it is expected that the taxfiler counts will be less than the census population. And insofar as the rate of taxfiling can vary between census tracts, the coverage can also be expected to vary for reasons other than geographical misallocations.

The analysis of coverage proved to be a means of identifying errors in the conversion file -- some postal codes were assigned to the wrong census tract. Attention was then focussed on two extreme coverage ratios -- 30 percent and 158 percent. (The remaining 82 coverage ratios ranged between 39 percent and 62 percent with a mean of 52.3 percent.)

In the case of the low total population coverage ratio, it was possible to determine that Census Tract 42 is heavily populated by elderly residents. After exploring several alternative explanations it was concluded that the high proportion of elderly residents reflected a higher non-filing rate than for other census tracts in the Calgary CMA.

The high total population coverage ratio (158 percent) was for Census Tract 43, a census tract for the central business district with a relatively low resident population. An in-depth review of the address information revealed that a large number of tax returns had a mailing address for accounting firms. Although this was to be expected, it is possible that future research will enable the development of procedures for reallocating some of these taxfilers to other addresses.

Finally, slightly lower total population coverage occured in the suburbs. This necessitated the second coverage ratio, taxfiler counts to population aged 15 years of age and older. With this rede-

finition of the population, and the elimination of the proportionately higher incidence of children in the suburbs, the population coverage ratios in the suburbs were much closer to the city average.

B. Ottawa-Hull

Basically, the same methodological steps were followed for Ottawa and Hull, although the analyses were not as extensive.

For the creation of the Ottawa CMA T1 file, all taxfilers with the locality code "15900" or the postal codes beginning with "K1" or "K2" were selected. For Hull, all taxfilers with the locality code "09300" or postal codes beginning with "J8" or "J9" were selected. Using these selection procedures, the Ottawa file contained 316,397 individual records while the Hull file contained 72,881 records. A summary of the results for Calgary, Ottawa and Hull is included as TABLE 5 for purposes of comparison.

The residual problems found in Ottawa and Hull were the same as those for Calgary (i.e., tax records without postal codes, incomplete conversion file and postal codes that could not be assigned to any census tracts).

TABLE 5: Summary of Results from Two Pilot Projects

	Calgary		Ottawa		Hull	
	Number	%	Number	%	Number	%
Assigned Census Tracts	246,145	87.60	290,013	91.66	65,548	89.94
No Census Tracts Assigned	34,836	12.40	26,354	8.34	7,333	10.06
Total T1 Records	280,981	100.00	316,397	100.00	72,881	100.00

Coverage ratios were also computed for the population of tax-filers in Ottawa and Hull. (Maps and coverage ratios for Ottawa and Hull are included in APPENDIX 3.) The total population coverage ratios for the 105 Ottawa census tracts varied from 44 percent to 82 percent, excluding several extreme values (see APPENDIX 3). The extreme values were found in Census Tracts 48, 54, 125, 126, 127, 141 and 151 to 200. Census Tracts 48 and 54 had very high coverage(i.e.,157 percent and 106 percent, respectively). Both of these census tracts are located in the central business district and again highlight the tax accountant problem of Calgary. For the fringe areas, there was low total population coverage, a phenomenon similar to Calgary. For example, reference to the Ottawa map in APPENDIX 3 indicates that Census Tract 125 (south-east), Census Tract 126 (south) and Census Tract 141 (west and south-west) are essentially low density urban fringe areas with rural mail delivery. In the case of Census Tract 141, the coverage ratio was computed to be one percent. There were still other residents of the Ottawa Regional Municipality of Ottawa-Carleton that were not included. Some of these census tracts are identified by place names other than those coded to Ottawa locality code and are not served by the Ottawa Post Office.

Similar situations were found in Hull for total population coverage (APPENDIX 3). After excluding five low coverage ratios (i.e., 0, 3, 4, 5 and 6 percent), the remaining total population coverage ratios ranged between 36 and 56 percent. Hull's suburban areas were also

characterized by low total population coverage ratios. The suburban ratios were improved by computing coverage ratios for the population 15 and over. But for both Ottawa and Hull, the coverage ratios for the 15 and over population were adversely affected by the rural route exclusion problem noted above. As a result, the fringe area coverage ratios were less susceptible to improvement than was the case for Calgary. In 1976 the central business district in Hull was disrupted by major construction of federal office buildings. As a consequence of this, overcoverage in the Hull central business district did not occur.

C. Summary of Census Tract Pilot Studies

Overall, and from a methodological perspective, the assignment of census tract codes to T1 records proved to be extremely promising. And in view of these results, the work will be expanded and extended to other CMA's.

In spite of the positive findings, three challenges remain to be addressed:

i. How can the high coverage ratios be reduced in the central business district of a CMA?

ii. How can the low coverage ratios in suburban areas be improved, especially in those CMA's with rural or suburban mail delivery?

iii. How can taxfilers using box numbers be allocated to census tracts?

V. CONCLUDING REMARKS

A. Summary of Findings

An overview has been provided in this paper of two projects undertaken by the Administrative Data Development Staff.

In the first project, locality codes and postal codes were converted into census division codes. It was concluded that the conversion of either a postal code or a locality code into a census division code produced almost identical results. This project did not, however, include an assessment of accuracy (an extensive coverage project is currently underway).

A second project was undertaken (divided into two pilot studies) to assess the feasibility of using postal codes to assign census tract codes to individual administrative records. For the pilot CMAs, two extreme cases emerged:

a. Central Business Districts. In CBDs, the number of taxfilers tended to exceed the resident population. This phenomenon was mainly attributed to individuals filing their tax returns through the offices of their accountants.

b. Rural/Suburban Areas. Low coverage occurred in the rural/suburban fringe areas. The low coverage attributed to suburban service and rural service -- taxfilers could not be assigned to their respective census tracts of residence.

In spite of the extreme cases, there was every indication that the data derived from census tract coded records provided patterns consistent with the general population. Therefore, the work proved to be sufficiently encouraging for census tract coding to be extended to additional CMAs.

Another objective of this paper was stated as being "a general assessment of the extent to which mailing addresses approximate place of residence". On the basis of the findings to date, the verdict has yet to be determined. But the potential, based on over two years of experience, can be summarized as follows:

a. Census Divisions. It is apparent that few coding misallocations occur in urban areas as a result of using the postal code, irrespective of whether the address is a mailing address or a residential address. In rural areas, however, considerable fuzziness occurs along CD boundaries since postal service areas do not respect these boundaries. Nevertheless, preliminary evidence indicates that these errors tend to cancel out.

b. Census Tracts. The coveargе ratios indicate considerable symmetry between taxfiler counts and census counts for census tracts, if the CBD and rural/suburban fringe extremes are ignored.

Although there is evidence to support the consistency of mailing and residential addresses, there are a number of exceptions. Therefore, it is likely that a more definitive assessment of the consistency of these addresses must be deferred until additional analysis is completed.

B. Some Proposed Solutions

Throughout this paper, a number of caveats and limitations have been identified with respect to the conversion of mailing addresses into census divisions and census tract codes. In general, the solutions can be summarized into three generic classes as follows:

a. improved address information,

b. more complete postal coding,

c. extended conversion files.

a. Address Information Improvements. It is conceptually possible to use alternative data sources to improve the address information of administrative records -- telephone directories, city directories (when available), and, perhaps, other administrative records systems. Furthermore, it is possible to conduct in-depth studies of those geographical areas with relatively weak address information, and to develop criteria for allocating individual taxfilers into some areas (e.g., rural/suburban fringe areas) and out of other areas (e.g., CBDs).

b. More Complete Postal Coding. Many mailing addresses do not contain postal codes (about 10 percent or 1.5 million 1979 T1 records). This does not represent a dilemma of major proportions. A software package has been developed experimentally and mailing addresses are being assigned postal codes when they are missing to decrease the current reliance on the locality code. Unfortunately, a universally applicable system has not been developed since address conventions vary between

geographical areas. Thus, the software is being developed iteratively as modifications are made for different areas. It is believed, however, that over time a comprehensive system will be developed for the automatic assignment of postal codes to individual records when the postal codes are missing.

c. Extended Conversion Files. The existing conversion files have numerous limitations. The Standards Division file had deficiencies in municipalities or CMAs that are census tract coded. The Geography Division has a very comprehensive Area Master File (i.e., AMF), but it does not contain postal codes. Furthermore, Statistics Canada does not currently possess a conversion file that adequately meets conversion needs in rural areas. Although these are currently identifiable limitations, a number of proposals exist that, if implemented, can reduce this array of limitations. (And as an aside, it seems fairly clear that additional conversion files are required for electoral districts, provincial planning areas, and an array of user-specified areas.)

C. Concluding Overview

In spite of the limitations noted above, it is clear that the mailing address information can be used effectively and efficiently to produce administrative social data for standard geographical areas.

And while there are some misallocations, and while all of the geographical inconsistencies have yet to be resolved, explained or even understood, the results are promising. Therefore, the future for the production of tabulations from administrative records for small areas is virtually assured. This is quite a different statement than could have been made even two years ago. At that time, an even tentative verdict had yet to be rendered -- it was believed then, that there would be no future for small area administrative social data if mailing addresses could not be used to delineate small areas. There is no longer any doubt that small areas can be defined using mailing address information.

Given the successes in using the mailing address, and given the similar results when administrative tax data were compared with 1976 Census of Population data, it is now clear that useful small area data can be produced annually. Therefore, the priorities have been changed to the production of experimental administrative tax data to enable the development and assessment of time series data, and to enable producers and users of small area data the opportunity to identify the reasonableness and potential uses of these data.

One of the priorities that has been defined lies in the domain of "tiny area" data development. Twenty-nine very small or "tiny" areas of Prince Edward Island have been defined and work has been underway for about six months to determine the feasibility, reliability and

usefulness of producing administrative tax data for these areas. To date the work has been promising, and data development efforts are continuing at a relatively slow pace because of the absolute necessity to preserve confidentiality, a task that varies inversely with the size of the geographical area. But there is no doubt that data for very small, user-defined areas can be produced. The results of the data produced for the 29 "tiny" areas of Prince Edward Island are a clear testament of the small area administrative data potential.

The overall implications of the development work is inescapable: the mailing address has every indication of being an acceptable surrogate for the conventional census/survey residential address. The mailing address continues to have some problems, but for a vast majority of Canadian taxfilers the mailing and residential addresses appear to be identical. Furthermore, the work conducted on the conversion of mailing address information to standard geographical areas has been exceedingly promising.

The Administrative Data Development Staff is now undertaking major development efforts to exploit administrative records for data that <u>may</u> prove useful to producers and users of social data. The data <u>may</u> be useful in the sense that the kinds of data produced, the areal configurations, and the tabulations largely represent examples of what can be done. The decisions on what will be produced, in what configurations, and for what geographical areas will hopefully be determined by the substantive contributions of the producers and users of social data.

Bibliography

1. Canada Post. Ontario Directory for Postal Code, Ottawa, March 1979/80.

2. Kopustas, N.J. Geography Project Postal Code - Locality Code Consistency, Ottawa, December 1979. (Unpublished paper).

3. Statistics Canada. 1976 Census of Canada Census Tracts, Population and Housing Characteristics, Calgary, Catalogue 95-801, Ottawa, November 1978.

4. Statistics Canada. 1976 Census of Canada, Census Tracts, Population and Housing Characteristics, Ottawa-Hull, Catalogue No.95-813, Ottawa, November 1978.

5. Statistics Canada. Standards Division, Postal Code - Geographical Classification Code Conversion File User's Guide, Ottawa, March 1979.

6. Statistics Canada. Standards Division, Postal Delivery Area - Geographical Classification Code Conversion Table, Ottawa, March 1979.

7. Statistics Canada. Standards Division, Changes to the Standard Geographical Classification for 1981, Ottawa, August 1980.

APPENDICES

APPENDIX 1 Examples of Combined Census Divisions

Locality Code	SGC Codes	Census Division Names
078	2411	Charlevoix-est
	2412	Charevoix-ouest
089	2401	Iles-de-la-Madelaine
	2402	Gaspé-est
	2403	Gaspé-ouest
091	2464	Ile-Jésus
	2465	Ile-de-Montréal
092	2478	Gatineau
	2479	Hull
100	2490	Lac-St.-Jean-ouest
	2493	Lac-St.-Jean-est
108	2405	Matapédia
	2406	Matane
114	2416	Montmorency No.2
	2417	Montmerency No.1
128	2497	Saguenay
	2499	Nouveau-Québec
143	2408	Rivière-du-Loup
	2409	Témiscouata
223	3552	Sudbury
	3553	Sudbury-Régional

APPENDIX 2

Examples of tabulations from Reports 1, 2, 3 and 4

Note: Report 1 contains examples of records with postal codes not found in the postal code SGC census division code table and, the taxfiler count for those postal codes.

Report 2 contains examples of locality codes not found in the locality code-SGC census division code table, their respective postal codes and the taxfiler count.

Report 3 contains examples of records where $CD_P = CD_L$ listed by locality code, taxfiler count and percentages.

Report 4 contains examples of records with $CD_P \neq CD_L$ (inconsistencies), their postal code, locality code and the taxfiler count.

Report 1

Postal Code	Taxfiler Count
BCN1S0	121
B0W3G0	32
B3K2X0	87
G0J3M0	23
G0W3A0	34
H0A1G0	56
J0B1Z0	55
J0B3K0	68
J0N3A0	66
J0X2B0	20
J0Z2M0	45
K0K2R0	84
K1A0K5	15
L0E1M0	65
L0H1N0	38
L0P1R0	95
L3T4T1	12
L3T4T3	17
M3K2F6	19
N0G3A0	38
S0E2A0	14
T0A1K0	13
T0K1Y0	22
T5J0K3	11
T6G2E1	20
V0T1C0	45
V0T1H0	32
V0T1P0	51
V0T1Z0	16
V6C3L7	11
X0E1H0	288

Report 2

Locality Code	Postal Code	Taxfiler Count
00056	A0A2K0	1
00063	A0A2M0	1
00077	A0A3X0	1
00509	A0K2E0	1
00509	A2H6B9	1
00513	A0J1S0	4
00916	A0J1T0	6
00916	A0G3Z0	1
00916	A0J1T0	1
00929	A0G4M0	1
00968	A0K1S0	3
00968	A0K3R0	1
01014	A0K4Y0	1
01014	A0K1H0	2
01190	A0K3L0	1
01700	B2G2L1	1
02000	B3L4M6	1
08414	J0B1L0	1
08914	G0E1B0	1
09696	J0K1C0	1
10028	G0W1K0	18
10028	J0N1J0	2
11132	J0J2B0	12
11911	G0A3M0	9
19512	P0B1C0	1
24021	L0R1H0	1
24318	R0E0K0	1

Report 3

Locality Code	Taxfiler Count	Percentage
01349	5	0.00
01351	84	0.00
01352	146	0.01
01354	6	0.00
01356	1	0.00
01357	109	0.01
01358	30	0.00
01359	183	0.01
01401	24	0.00
01404	1	0.00
01406	1	0.00
01403	29	0.00
01409	20	0.00
01412	37	0.00
01414	51	0.00
01418	18	0.00
01419	100	0.01
01421	64	0.00
01424	131	0.01
01426	204	0.01
01428	11	0.00
01429	18	0.00
01431	10	0.00
01432	15	0.00
01434	27	0.00
01436	25	0.00
01441	121	0.01
01442	6	0.00
01444	11	0.00
01446	42	0.00
01449	14	0.00
01452	70	0.00
01454	4	0.00
01456	1	0.00
01458	1	0.00
01459	110	0.01
01500	1703	0.10
01601	52	0.00
01604	294	0.01
01603	120	0.01
01609	154	0.01
01612	194	0.01
01614	101	0.00
01616	343	0.02
01701	79	0.00
01702	336	0.02
01705	420	0.02
01801	60	0.00
01804	191	0.01
01809	1	0.00
01816	106	0.01
01832	2	0.00
01900	6272	0.37
02001	246	0.01
02002	118	0.01

Report 4

CD_P	CD_L	LC	PC	Taxfiler Count
1009	1008	00974	A0K1Z0	5
1009	1008	00974	ACK3N0	1
1009	1008	00974	ACK4D0	1
1009	1008	00976	A0K4S0	1
1009	1010	01112	A0K1W0	2
1009	1010	01132	A0K4N0	1
1009	9999	80000	A0K1M0	1
1010	1001	00036	ACK3P0	20
1010	1001	00036	A2V1R5	1
1010	1001	00053	A0K3X0	1
1010	1001	00100	A0K3K0	1
1010	1001	00100	A0P1J0	1
1010	1001	00100	A0R1A0	2
1010	1001	00100	A0R1B0	1
1010	1004	00419	A0R1D0	1
1010	1007	00031	A0K5Y0	1
1010	1008	00946	A0K1E0	1
1010	1009	01001	ACK2L0	1
1010	1009	01001	ACK2S0	1
1010	1009	01010	A0K2P0	1
1010	1209	02612	A2V1F6	1
1010	4006	42800	A0R1S0	1
1010	7001	75799	A2V1Y3	1
1010	9999	80000	A0K5H0	1
1101	1102	01401	C0A1W0	1
1101	1102	01404	C0A1R0	1
1101	1102	01419	C0A1W0	1
1101	1102	01426	C0A1V0	1
1101	1102	01431	C0A1V0	1
1101	1102	01432	C0A1R0	1

APPENDIX 3

Coverage ratios were calculated as follows:

1) 'Population Total'

Total number of taxfilers as a percentage of the total census population.

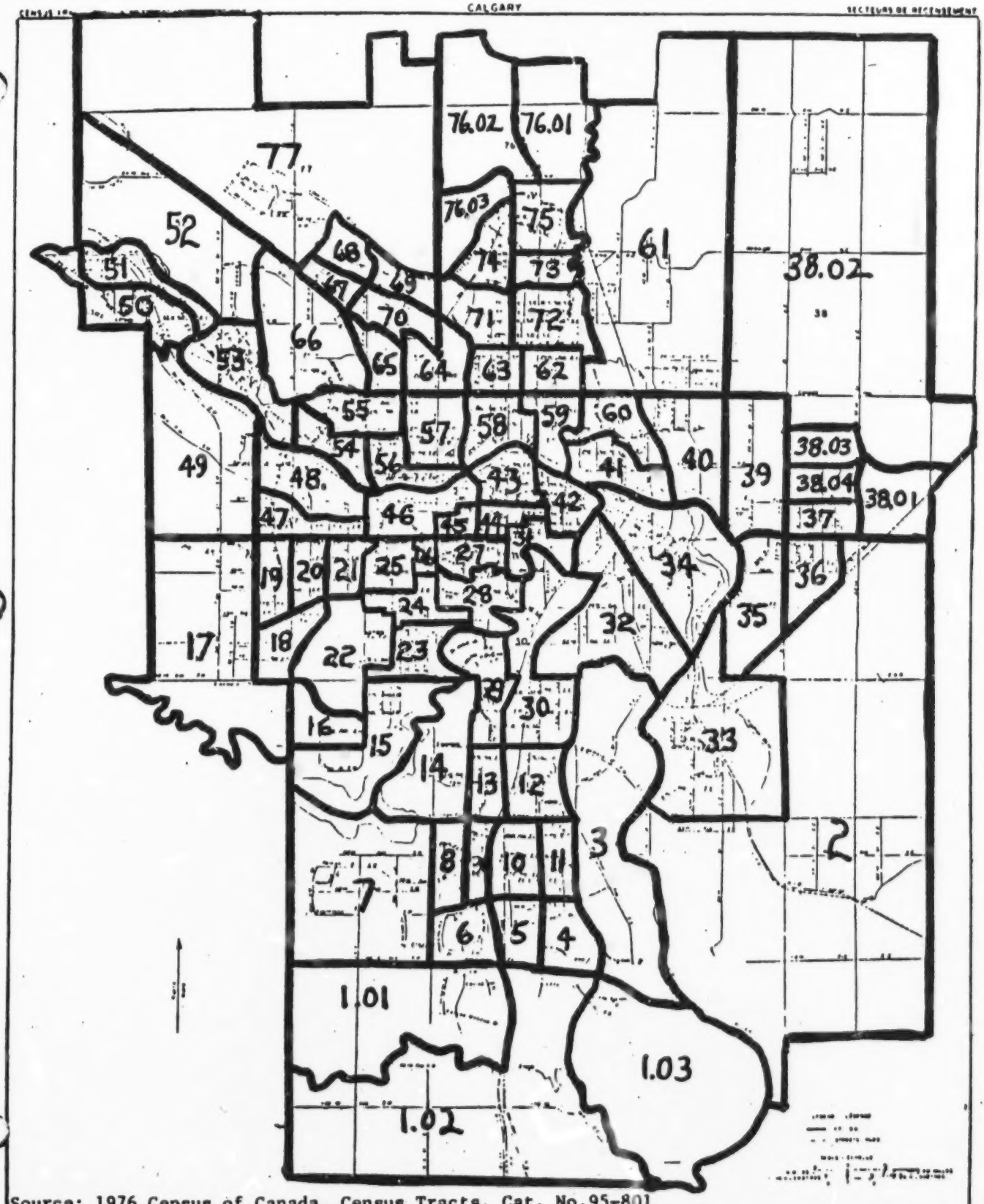
e.g., In Calgary Census Tract 1.01 there were 3,343 taxfilers and a population of 6,773 for a coverage ratio of 49%.

2) 'Population 15 and over'

Total number of taxfilers as a percentage of the census population 15 years of age and over.

e.g., In Calgary Census Tract 1.01 there were 3,343 taxfilers and a population of 15 years and over of 4,325 for a coverage ratio of 77%.

MAP 1



Source: 1976 Census of Canada, Census Tracts, Cat. No.95-801

Table (i)

Taxfiler Coverage for Calgary Census Tracts

Census Tract Code	Population Total (%)	Population 15 and over (%)	Census Tract Code	Population Total (%)	Population 15 and over (%)
101	49	77	3200	52	67
102	45	68	3300	52	76
103	47	74	3400	51	63
200	39	51	3500	43	67
300	47	71	3600	44	68
400	46	67	3700	50	73
500	52	75	3801	42	68
600	49	73	3802	55	83
700	47	71	3803	47	79
800	56	76	3804	45	69
900	56	72	3900	48	69
1000	42	59	4000	46	68
1100	50	72	4100	46	55
1200	54	74	4200	30	34
1300	58	70	4300	158	169
1400	53	65	4400	62	64
1500	51	69	4500	54	60
1600	49	69	4600	57	64
1700	45	64	4700	53	68
1800	48	67	4800	52	65
1900	55	71	4900	53	70
2000	57	71	5000	55	79
2100	58	69	5100	49	67
2200	44	66	5200	50	74
2300	56	70	5300	47	62
2400	55	66	5400	58	70
2500	58	68	5500	59	70
2600	54	60	5600	59	70
2700	59	66	5700	51	60
2800	54	67	5800	60	69
2900	60	71	5900	57	66
3000	55	67	6000	52	66
3100	54	57	6100	48	68

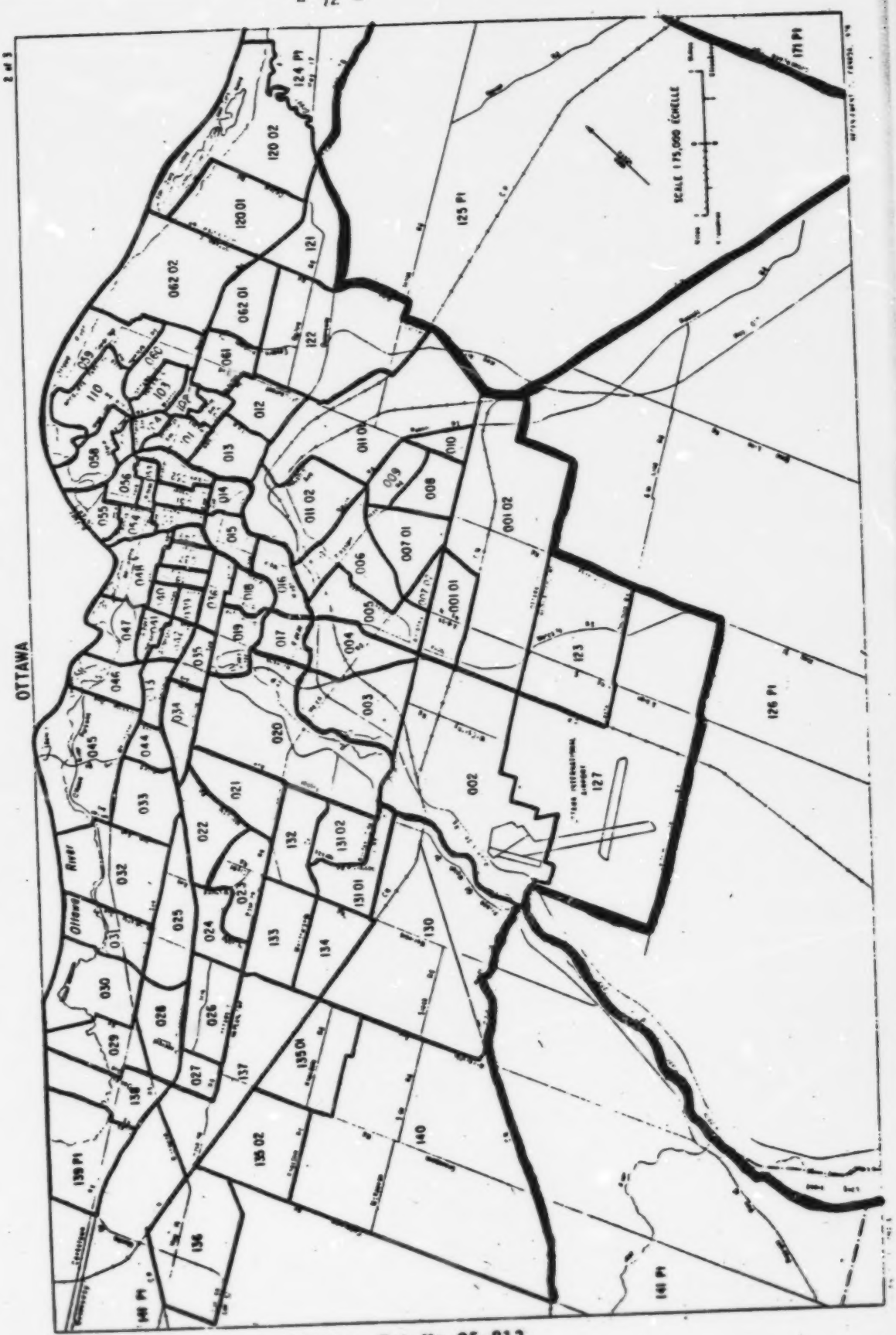
Note: Census Tract codes in the above table correspond to census tract codes on Map 1 in the following manner: 101=1.01, 6200=62, 7603=76.03.

Table (i) Continued

Census Tract Code	Population Total(%)	Population 15 and over(%)
6200	58	69
6300	58	67
6400	61	73
6500	61	72
6600	55	74
6700	56	71
6800	53	74
6900	57	73
7000	58	74
7100	56	71
7200	54	68
7300	52	66
7400	61	76
7500	49	69
7601	50	79
7602	46	70
7603	55	78
7700	47	70
Total	52.3	70.3

Note: Census Tract codes in the above table correspond to census tract codes in Map 1 in the following manner: 101=1.01, 6200=62, 7603=76.03.

MAP 2



Source: 1976 Census of Canada, Census Tracts, Cat.No.95-813.

Table (ii)

Taxfiler Coverage for Ottawa Census Tracts

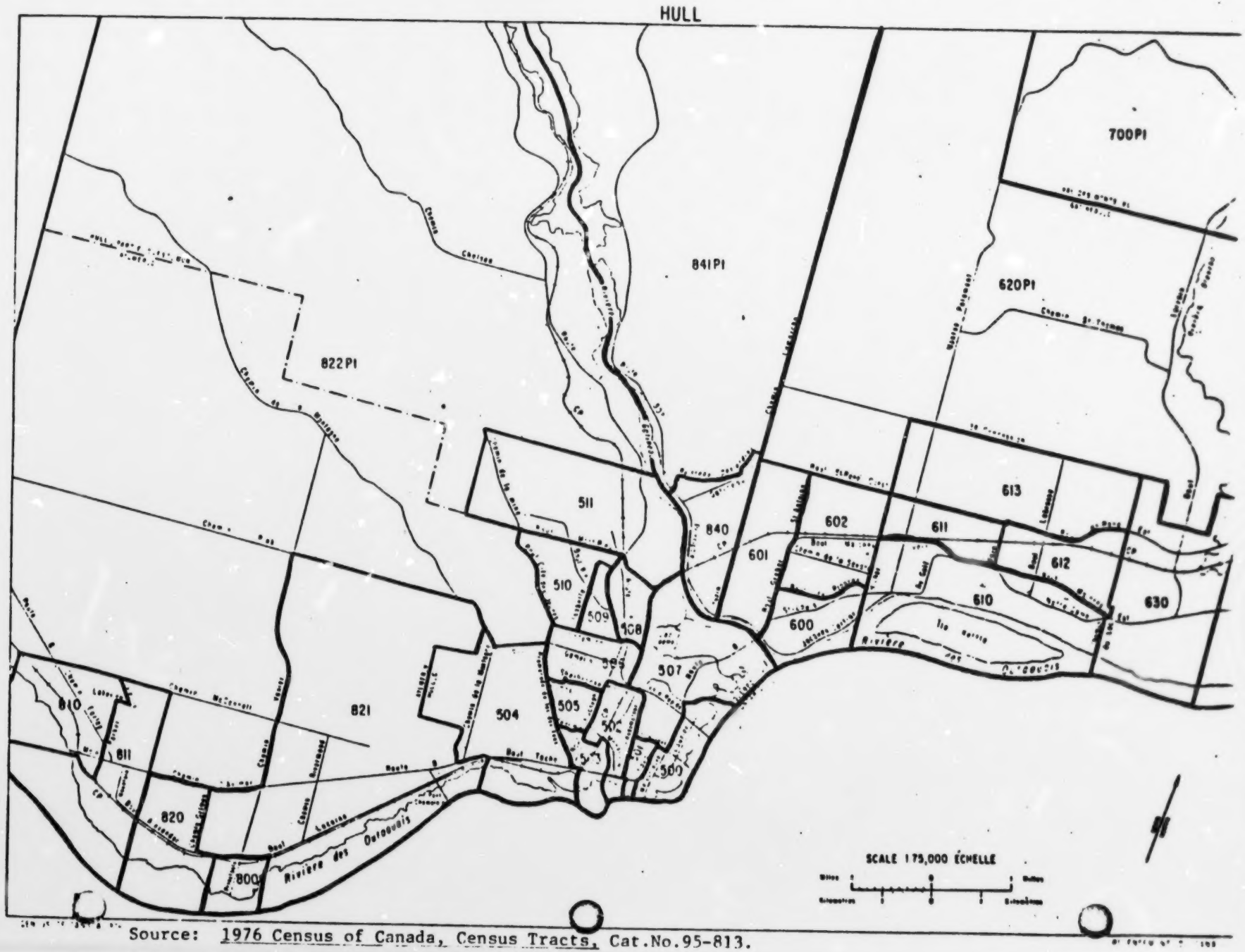
Census Tract Code	Population Total(Z)	Population 15 and over(Z)	Census Tract Code	Population Total(Z)	Population 15 and over(Z)
101	48	69	3100	75	84
102	71	101	3200	68	78
200	60	82	3300	63	76
300	69	83	3400	61	77
400	61	76	3500	62	74
500	63	75	3600	59	70
600	60	72	3700	74	77
701	51	67	3800	82	86
702	57	72	3900	60	75
800	61	78	4000	61	72
900	64	78	4100	50	62
1000	49	66	4200	57	72
1101	58	76	4300	62	77
1102	63	73	4400	64	77
1200	51	73	4500	74	88
1300	61	77	4600	59	74
1400	80	117	4700	82	77
1500	47	55	4800	157	169
1600	66	77	4900	55	60
1700	58	71	5000	80	85
1800	65	76	5100	69	75
1900	68	83	5200	69	74
2000	61	79	5300	63	71
2100	57	73	5400	106	122
2200	59	76	5500	53	59
2300	62	77	5600	56	69
2400	60	74	5700	54	64
2500	61	73	5800	51	59
2600	59	79	5900	55	73
2700	51	71	6000	60	69
2800	51	73	6100	57	74
2900	49	67	6201	57	79
3000	57	69	6202	53	74

Table (ii) Continued

Census Tract Code	Population Total(%)	Population 15 and over(%)
10000	60	74
10100	66	78
10200	56	72
10300	57	73
10400	57	72
11000	46	58
12001	48	73
12002	44	70
12100	54	75
12200	55	80
12300	49	68
12400	56	69
12500	36	55
12600	-	-
12700	31	47
13000	54	78
13101	54	77
13102	56	73
13200	57	73
13300	50	67
13400	50	71
13501	52	77
13502	47	73
13600	63	89
13700	45	65
13800	57	75
13900	47	68
14000	50	82
14100	1	1
15000	51	-
15100	-	-
16000	35	53
17000	25	40
17100	-	-
18000	-	-
18100	-	-
19000	-	-
19100	-	-
20000	-	-
Total	55.6	72.6

Note: -(dash) indicates 0(zero) percent.

Note: Census Tract codes in the above table correspond to census tract codes on Map 2 in the following manner: 101=001.01, 2000=020, 6202=062.02.



Source: 1976 Census of Canada, Census Tracts, Cat.No.95-813.

Table (iii)

Taxfiler Coverage for Hull Census Tracts

Census Tract Code	Population Total(Z)	Population 15 and over(Z)
500	43	53
501	45	57
502	43	55
503	45	58
504	50	62
505	48	62
506	56	69
507	38	53
508	48	60
509	49	63
510	50	65
511	42	57
600	39	5
601	36	55
602	41	60
610	43	58
611	41	62
612	43	61
613	42	64
620	6	9
630	39	56
700	-	-
800	23	32
810	37	54
811	39	59
820	43	63
821	36	48
822	3	5
840	5	8
841	4	5
Total	38.1	53.5

Note: - (dash) indicates 0(zero) percent.